Laboratory Phonology
Journal of the Association for Laboratory Phonology

Roettger, T. B. 2019 Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 1, pp. 1–27. DOI: https://doi.org/10.5334/labphon.147

JOURNAL ARTICLE

Researcher degrees of freedom in phonetic research

Timo B. Roettger
Department of Linguistics, Northwestern University, Evanston, IL, US
timo.b.roettger@gmail.com

The results of published research critically depend on methodological decisions that have been made during data analysis. These so-called 'researcher degrees of freedom' (Simmons, Nelson, & Simonsohn, 2011) can affect the results and the conclusions researchers draw from it. It is argued that phonetic research faces a large number of researcher degrees of freedom due to its scientific object—speech—being inherently multidimensional and exhibiting complex interactions between multiple covariates. A Type-I error simulation is presented that demonstrates the severe inflation of false positives when exploring researcher degrees of freedom. It is argued that combined with common cognitive fallacies, exploitation of researcher degrees of freedom introduces strong bias and poses a serious challenge to quantitative phonetics as an empirical science. This paper discusses potential remedies for this problem including adjusting the threshold for significance; drawing a clear line between confirmatory and exploratory analyses via preregistration; open, honest, and transparent practices in communicating data analytical decisions; and direct replications.

Keywords: false positive; methodology; preregistration; replication; reproducibility; speech production; statistical analysis

1. Introduction

Data analysis—that is, the path we chose from the raw data to the results section of a paper—is a complex process. We can look at data from different angles and each way to look at them may lead to different methodological and analytical choices. These potential choices are collectively referred to as researcher degrees of freedom (Simmons et al., 2011). Said choices, however, are often not specified in advance but are made in an ad hoc fashion, after having explored several aspects of the data and analytical choices. In other words, they are data-contingent rather than motivated on independent, subject-matter grounds. As is argued below, exploiting researcher degrees of freedom is often not an intentional process. However, it needs to be addressed because exploiting researcher degrees of freedom, intentionally or not, increases the chances of finding a false positive, i.e., finding a pattern that is incorrectly interpreted as rejecting the null hypothesis. This problem is shared by all quantitative scientific fields (Gelman & Loken, 2014; Simmons et al., 2011; Wicherts et al., 2016), but has not been extensively discussed for the specific characteristics of phonetic data analyses.

In this paper, I will argue that analyses in quantitative phonetics face a high number of researcher degrees of freedom due to the inherent multidimensionality of speech behavior, which is the outcome of a complex interaction between different functional layers. This article will discuss relevant researcher degrees of freedom in quantitative phonetic research, reasons as to why exploiting researcher degrees of freedom is potentially harmful for phonetics as a cumulative empirical endeavor, and possible remedies to these issues.

The remainder of this paper is organized as follows: In Section 2, I will review the concept of researcher degrees of freedom and how they can lead to misinterpretations

in combination with certain cognitive biases. In Section 3, I will argue that researcher degrees of freedom are particularly prevalent in quantitative phonetics, focusing on the analysis of speech production data. In Section 4, I will present a simple simulation, demonstrating that chance processes lead to a large inflation of false positives when we exploit common researcher degrees of freedom. In Section 5, I will discuss possible ways to reduce the probability of false positives due to researcher degrees of freedom, discussing adjustments of the alpha level (Section 5.1), stressing the importance of a more rigorous distinction between confirmatory and exploratory analyses (Section 5.2), preregistrations and registered reports (Section 5.3), transparent reporting (Section 5.4), and direct replications (Section 5.5).

2. Researcher degrees of freedom

Every data analysis is characterized by a multitude of decisions that can affect its outcome and, in turn, the conclusions we draw from it (Gelman & Loken, 2014; Simmons et al., 2011; Wicherts et al., 2016). Among the decisions that need to be made during the process of learning from data are the following: What do we measure? What predictors and what mediators do we include? What type of statistical models do we use?

There are many choices to make and most of them can have an influence on the results that are obtained. Often these decisions are not made prior to data collection: Instead, we often explore the data and possible analytical choices to eventually settle on one 'reasonable' analysis plan which, ideally, yields a statistically convincing result. This paper argues that our statistical results are strongly affected by the number of hidden analyses performed.

In most scientific papers, statistical inference is drawn by means of null hypothesis-significance-testing (NHST, Gigerenzer, Krauss, & Vitouch, 2004; Lindquist, 1940). Because NHST is by a large margin the most common inferential framework used in quantitative phonetics, the present discussion is conceived with NHST in mind. Traditional NHST performs inference by assuming that the null hypothesis is true in the population of interest. For concreteness, assume we are conducting research on an isolated undocumented language and investigating the phonetic instantiation of phonological contrasts. We suspect the language has word stress, i.e., certain syllables are phonologically 'stronger' than other syllables within a word. We are interested in how word stress is phonetically manifested and pose the following hypothesis:

(1) There is a phonetic difference between stressed and unstressed syllables.

In NHST, we compute the probability of observing a result at least as extreme as a test statistic (e.g., *t*-value), assuming that the null hypothesis is true (the *p*-value). In our concrete example, the *p*-value tells us the probability of observing our data or more extreme data, if there was no difference between stressed and unstressed syllables (null hypothesis). Receiving a *p*-value below a certain threshold (commonly 0.05) is then interpreted as evidence to claim that the probability of the data, if the null hypothesis was in fact true (no difference between stressed and unstressed syllables), is sufficiently low. This is henceforth considered a positive result.

One common error within this framework that can occur is a false positive, i.e., incorrectly rejecting a null hypothesis (Type I error).[1] When undetected, false positives

[1] There are other errors that can happen and that are important to discuss. Most closely related to the present discussion are Type II errors (i.e., false negatives, Thomas et al., 1985), Type M(agnitude), and Type S(ign) errors (Gelman & Carlin, 2014). Within quantitative linguistics, these errors have recently been discussed by, for example, Kirby and Sonderegger (2018), Nicenboim, Roettger, and Vasishth (2018a), Nicenboim and Vasishth (2016), and Vasishth and Nicenboim (2016).

can have far reaching consequences, often leading to theoretical claims that may misguide future research (Smaldino & McElreath, 2016). These errors can be persistent through time because our publication system neither incentivizes publishing null results nor direct replication attempts, biasing the scientific record toward novel positive findings. As a result, there may be a large number of null results in the 'file drawer' that will never see the light of day (e.g., Sterling, 1959).

Within the NHST framework, any difference between conditions that yields a *p*-value below 0.05 is, in practice, considered sufficient to reject the null hypothesis and to claim that there is a difference. However, these tests have a natural false positive rate, i.e., given a *p*-value of 0.05, there is a 5% probability that our data accidentally suggest that the null hypothesis can be refuted.

Coming back to our hypothetical example, if, for example, we decide to measure only a single phonetic parameter (e.g., vowel duration) to test the hypothesis in (1), 5% would be the base rate of false positives, given a *p*-value of 0.05 (and assuming that the null hypothesis is true). However, this situation changes if we measure more than one parameter. For example, we could test, say, vowel duration, average intensity, and average f_0 (all common phonetic correlates of word stress, e.g., Gordon & Roettger, 2017), amounting to three null hypothesis significance tests. One of these analyses may yield a *p*-value of 0.05 or lower. We might proceed to write a paper based on this significant finding in which we argue that stressed and unstressed syllables are phonetically different in the language under investigation.

This procedure increases the chances of finding a false positive. If n independent comparisons are performed, the false positive rate would be $1-(1-0.05)^n$ instead of 0.05. Three tests, for example, will produce a false positive rate of approximately 14% (i.e., $1-0.95 * 0.95 * 0.95 = 1-0.857 = 0.143$). Why is that? Assuming we could get a significant result with a *p*-value of 0.05 by chance in 5% of cases, the more often we look at random samples, the more often we will accidentally find a significant result (e.g., Tukey, 1953).

This reasoning can be applied to all researcher degrees of freedom. With every analytical decision, with every forking path in the analytical labyrinth, with every researcher degree of freedom, we increase the likelihood of finding significant results due to chance. In other words, the more we look, explore, and dredge the data, the greater the likelihood of finding a significant result. Exploiting researcher degrees of freedom until significance is reached has been called out as harmful practices for scientific progress (John, Loewenstein, & Prelec, 2012). Two often discussed instances of such harmful practices are HARKing (Hypothesizing After Results are Known, e.g., Kerr, 1998) and *p*-hacking (e.g., Simmons et al., 2011). HARKing refers to the practice of presenting relationships that have been obtained after data collection as if they were hypothesized in advance. *P*-hacking refers to the practice of hunting for significant results in order to ultimately report these results as if confirming the planned analysis. While such exploitations of researcher degrees of freedom are certainly harmful to the scientific record, there are good reasons to believe that they are, more often than not, unintentional.

People are prone to cognitive biases. Our cognitive system craves coherency and we are prone to seeing patterns in randomness (apophenia, Brugger, 2001); we weigh evidence in favor of our preconceptions more strongly than evidence that challenges our established views (confirmation bias, Nickerson, 1998); we perceive events as being plausible and predictable after they have occurred (hindsight bias, Fischhoff, 1975).[2] Scientists are no exception. For example, Bakker and Wicherts (2011) analyzed statistical errors in over

[2] See Greenland (2017) for a discussion of cognitive biases that are more specific to statistical analyses.

250 psychology papers. They found that more than 90% of the mistakes were in favor of the researchers' expectations, making a non-significant finding significant. Fugelsang, Stein, Green, and Dunbar (2004) investigated how scientists evaluate data that are either consistent or inconsistent with prior expectations. They showed that when researchers are faced with results that disconfirm their expectations, they are likely to blame the methodology while results that confirmed their expectations were rarely critically evaluated.

We work in a system that incentivizes positive results more than negative results (John et al., 2012), so we have the natural desire to find a positive result in order to publish our findings. A large body of research suggests that when we are faced with multiple decisions, we may end up convincing ourselves that the decision with the most rewarding outcome is the most justified one (e.g., Dawson, Gilovich, & Regan 2002; Hastorf & Cantril, 1954). In light of the dynamic interplay of cognitive biases and our current incentive structure in academia, having many analytical choices may lead us to unintentionally exploit these choices during data analysis. This can inflate the number of false positives.

3. The garden of forking paths in quantitative phonetics

Quantitative phonetics is no exception to the issues discussed above, and may in fact be particularly at risk because its very scientific object offers a considerable number of perspectives and decisions along the data analysis path. The next section will discuss researcher degrees of freedom in phonetics, and will, for exposition purposes, focus on speech production research. It turns out that the type of data we are collecting, i.e., acoustic or articulatory data, opens up many different forking paths (for a more discipline-neutral assessment of researcher degrees of freedom, see Wicherts et al., 2016). I discuss the following four sets of decisions (see **Figures 1** and **2**): choosing phonetic parameters (Section 3.1), operationalizing chosen parameters (Section 3.2), discarding data (Section 3.3), and choosing additional independent variables (Section 3.4). These distinctions are made for convenience and I acknowledge that there are no clear boundaries between these four sets. They highly overlap and inform each other to different degrees.

3.1. Choosing phonetic parameters

When conducting a study on speech production, the first important analytical decision to test a hypothesis is the question of operationalization, i.e., how to measure the phenomenon of interest. For example, how do we measure whether two sounds are phonetically identical, whether one syllable in the word is more prominent than others, or whether two discourse functions are produced with different prosodic patterns? In other words, how do we quantitatively capture relevant features of speech?

Speech categories are inherently multidimensional and vary through time. The acoustic parameters for one category are usually asynchronous, i.e., appear at different points of time in the unfolding signal and overlap with parameters for other categories (e.g., Jongman, Wayland, & Wong, 2000; Lisker, 1986; Summerfield, 1981; Winter, 2014). For example, the distinction between voiced and voiceless stops in English can be manifested by many different acoustic parameters such as voice onset time (e.g., Lisker & Abramson, 1963), formant transitions (e.g., Benkí, 2001), pitch in the following vowel (e.g., Haggard et al., 1970), the duration of the preceding vowel (e.g., Raphael, 1972), the duration of the closure (e.g., Lisker, 1957), as well as spectral differences within the stop release (e.g., Repp, 1979). Even temporally dislocated acoustic parameters correlate with voicing. For example, in the words led versus let, voicing correlates can be found in the acoustic

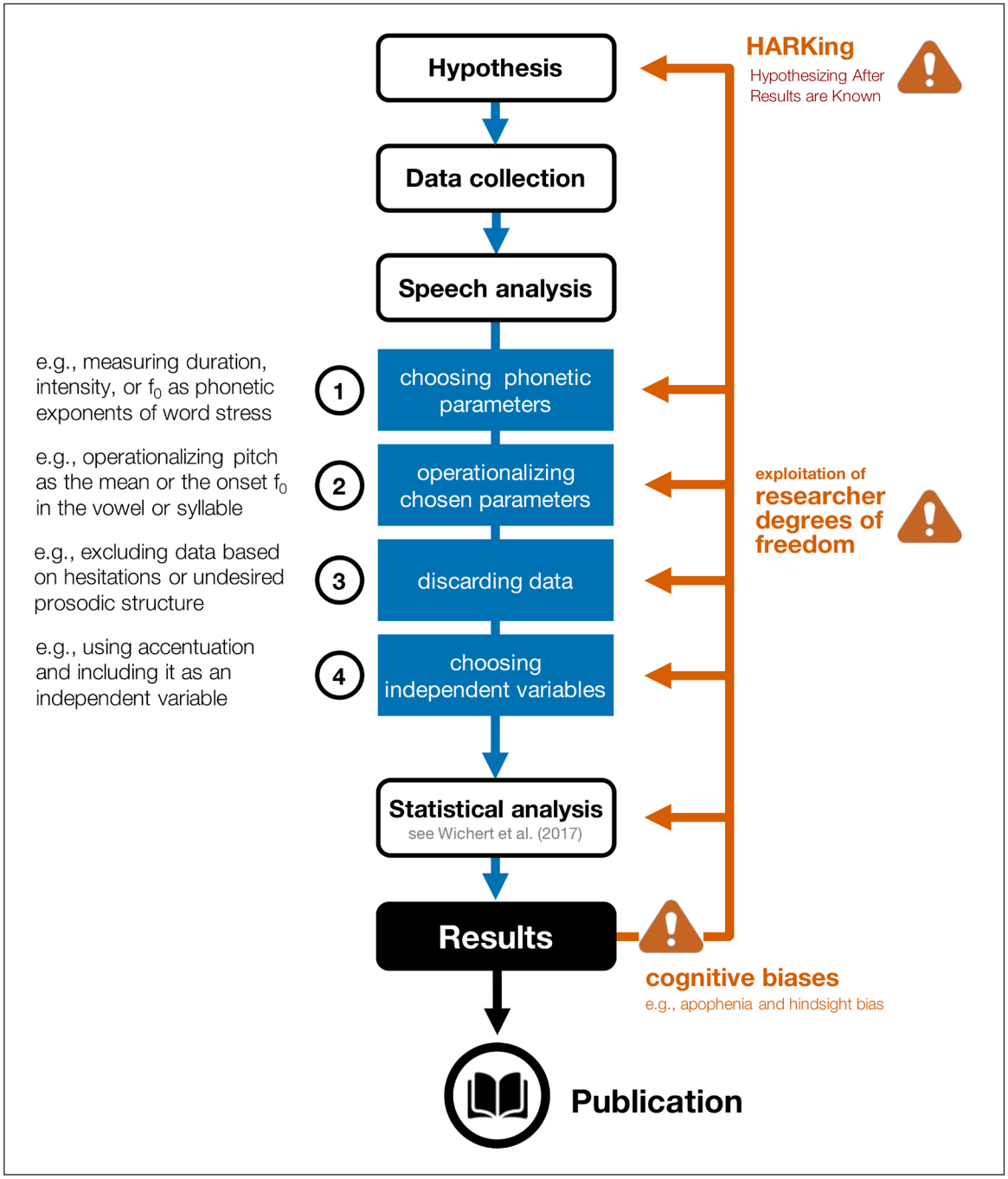

Figure 1: Schematic depiction of decision procedures during data analysis that can lead to an increased false positive rate. Along the first analysis pipeline (blue), decisions are made as to what phonetic parameters are measured (Section 3.1), how they are operationalized (Section 3.2), what data are kept and what data are discarded (Section 3.3), and what additional independent variables are measured (Section 3.4). The results are statistically analyzed (which comes with its own set of researcher degrees of freedom, see Wicherts et al., 2016) and interpreted. If the results are as expected and/or desired, the study will be published. If not, cognitive biases facilitate reassessments of earlier analytical choices (red arrow) (or a reformulation of hypotheses, i.e., HARKing), increasing the false positive rate.

manifestation of the initial /l/ of the word (Hawkins & Nguyen, 2004). These acoustic aspects correlate with their own set of articulatory configurations, determined by a complex interplay of different supralaryngeal and laryngeal gestures, coordinated with each other in intricate ways.

The multiplicity of phonetic cues grows exponentially if we look at larger temporal windows as is the case for suprasegmental aspects of speech. Studies investigating acoustic correlates of word stress, for example, have been using many different measurements including temporal characteristics (duration of certain segments or subphonemic intervals), spectral characteristics (intensity measures, formants, and spectral tilt), and measurements related to fundamental frequency (f_0) (Gordon & Roettger, 2017).

Looking at even larger domains, the prosodic expression of pragmatic functions can be expressed by a variety of structurally different acoustic cues which can be distributed throughout the whole utterance. Discourse functions are systematically expressed by multiple tonal events differing in their position, shape, and alignment (e.g., Niebuhr, D'Imperio, Gili Fivela, & Cangemi, 2011). They can also be expressed by global or local pitch scaling, as well as acoustic information within the temporal or spectral domain (e.g., Cangemi, 2015; Ritter & Roettger, 2014; van Heuven & van Zanten, 2005).[3] All of these phonetic parameters are potential manifestations of a communicative function of interest and therefore researcher degrees of freedom.

If we ask questions such as "is a stressed syllable phonetically different from an unstressed syllable?", any single measure has the potential to reject the corresponding null hypothesis (there is no difference). But which measurement should we pick? There are often many potential phonetic correlates for the relevant phonetic difference under scrutiny. Looking at more than one measurement seems to be reasonable. However, looking at many phonetic measurements to test a single global hypothesis increases the probability of finding a false positive. If we were to test 20 measurements of the speech signal repeatedly, on average one of these tests will, by mere chance, result in a spurious significant result (at a 0.05 alpha level). We obviously have justified preconceptions about which phonetic parameters may be good candidates for certain functional categories, informed by a long list of references. However, while these preconceptions can certainly help us make theoretically-informed decisions, they may also bear the risk for *ad hoc* justifications of analytical choices that happen *after* having explored researcher degrees of freedom.

One may further object that phonetic parameters are not independent of each other, i.e., many measurements covary systematically. Such covariation between multiple acoustic parameters can, for instance, result from the same underlying articulatory configurations. For example, VOT and onset f_0, the fundamental frequency at the onset of the vowel following the stop, are systematically covarying across languages, which has been argued to be a biomechanical consequence of articulatory and/or aerodynamic configurations (Hombert, Ohala, & Ewan, 1979). Löfqvist, Baer, McGarr, and Story (1989) showed that when producing voiceless consonants, speakers exhibit higher levels of activity in the cricothyroid muscle and in turn greater vocal fold tension. Greater vocal fold tension is associated with higher rates of vocal fold vibration leading to increased f_0. Since VOT and onset f_0 are presumably originating from the same articulatory configuration, one could argue that we do not have to correct for multiple testing when measuring these two acoustic parameters. However, as will be shown in Section 4, correlated measures lead to

[3] Speech is not an isolated channel of communication; it cooccurs in rich interactional contexts. Beyond acoustic and articulatory parameters, spoken communication is accompanied by non-verbal modalities such as body posture, eye gaze direction, head movement, and facial expressions, all of which have been shown to contribute to comprehension and may thus be considered relevant parameters to measure (Cummins, 2012; Latif, Barbosa, Vatiokiotis-Bateson, Castelhano, & Munhall, 2014; Prieto, Puglesi, Borràs-Comes, Arroyo, & Blat, 2015; Rochet-Capellan, Laboissière, Galván, & Schwartz, 2008; Yehia, Rubin, & Vatiokiotis-Bateson, 1998).

false positive inflation rates that are nearly as high as in independent multiple tests (see also von der Malsburg & Angele, 2017).

3.2. Operationalizing chosen parameters

The garden of forking paths is not restricted to choosing phonetic parameters. There are many different ways to operationalize the dimensions of speech that we have chosen. For example, when we want to extract specific acoustic parameters of a particular interval of the speech signal, we need to operationalize how to decide on a given interval. We usually have objective annotation procedures and clear guidelines that we agree on prior to the annotation process, but these decisions have to be considered researcher degrees of freedom and can potentially be revised after having seen the results of the statistical analysis.

Irrespective of the actual annotation, we can look at different acoustic domains. For example, a particular acoustic parameter such as duration or pitch can be operationalized differently with respect to its domain and the way it is assessed: In their survey of over a hundred acoustic studies on word stress correlates, Gordon and Roettger (2017) encountered dozens of different approaches how to quantitatively operationalize f_0, intensity and spectral tilt as correlates of word stress. Some studies took the whole syllable as a domain, others targeted the mora, the rhyme, the coda, or individual segments. Specific measurements for f_0 and intensity included the mean, the minimum, the maximum, or even the standard deviation over a certain domain. Alternative measures included the value at the midpoint of the domain, at the onset and offset of the domain, or the slope between onset and offset. The measurement of spectral tilt was also variegated. Some studies measured relative intensity of different frequency bands where the choice of frequency bands varied considerably across studies. Yet other studies measured relative intensity of the first two harmonics.

Another example of variable operationalizations can be found in time-series data such as pitch curves, formant trajectories, or articulatory movements. These time-series data have been analyzed as sequences of static landmarks ('magic moments,' Vatikiotis-Bateson, Barbosa, & Best, 2014), with large differences across studies regarding the identity and the operationalization of these landmarks. For example, articulatory studies looking at intragestural coordination commonly measure gestural onsets, displacement metrics, peak velocity, the onset and offset of gestural plateaus, or stiffness (i.e., relating peak velocity to displacement). Alternatively, time-series data can be analyzed holistically as continuous trajectories, differing with regard to the degrees of smoothing applied (e.g., Wieling, 2018).

In multidimensional data sets such as acoustic or articulatory data, there may be thousands of sensible analysis pathways. Beyond that, there are many different ways to process these raw measurements with regard to relevant time windows and spectral regions of interest; there are many possibilities of transforming or normalizing the raw data or smoothing and interpolating trajectories.

3.3. Discarding data

Setting aside the multidimensional nature of speech and assuming that we actually have *a priori* decided on what phonetic parameters to measure (see Section 3.1) and how to measure and process them (see Section 3.2), we are now faced with additional choices. Segmenting the acoustic signal may be difficult due to undesired speaker behavior, e.g., hesitations, disfluencies, or mispronunciations. There may be issues related to the quality of the recording such as background noise, signal interference, technical malfunctions, or interruptive external events (something that happens quite frequently during field work).

Another aspect of the data that may strike us as problematic when we extract phonetic information are other linguistic factors that could interfere with our research question. Speech consists of multiple information channels including acoustic parameters that distinguish words from each other and acoustic parameters that structure prosodic constituents into rhythmic units, structure utterances into meaningful units, signal discourse relations, or deliver indexical information about the social context. Dependent on what we are interested in, the variable use of these information channels may interfere with our research question. For example, a speaker may produce utterances that differ in their phrase-level prosodic make-up. In controlled production studies, speakers often have to produce a very restricted set of sentences. Speakers may for whatever reason (boredom, fatigue, etc.) insert prosodic boundaries or alter the information structure of an utterance, which, in turn, may drastically affect the phonetic form of other parts of the signal. For example, segments of accented words have been shown to be phonetically enhanced, making them longer, louder, and more clearly articulated (e.g., Cho & Keating, 2009; Harrington, Fletcher, & Beckman, 2000). Related to this point, speakers may use different voice qualities, some of which will make the acoustic extraction of certain parameters difficult. For example, if we were interested in extracting f_0, parts of the signal that are produced with a creaky voice may not be suitable; or if we were interested in spectral properties of the segments, parts of the signal produced in falsetto may not be suitable.

It is reasonable to 'clean' the data and remove data points that we consider as not desirable, i.e., productions that diverge from the majority of productions (e.g., unexpected phrasing, hesitations, laughter, etc.). These undesired productions may interfere with our research hypothesis and may mask the signal, i.e., make it less likely to find systematic patterns. We can exclude these data points in a list-wise fashion, i.e., subjects who exhibit a certain amount of undesired behavior (set by a certain threshold) may be entirely excluded. Alternatively, we can exclude trials on a pair-wise level across levels of a predictor. We can also simply exclude problematic tokens on a trial-by-trial basis. All of these choices change our final data set and, thus, may affect the overall result of our analysis.

3.4. Choosing independent variables

Often, subsetting the data or exclusion of whole clusters of data can have impactful consequences, as we would discard a large amount of our collected data. Instead of discarding data due to unexpected covariates, we can add these covariates as independent variables to our statistical model. In our example, we could include the factor of accentuation as an interaction term into our analysis to see whether the investigated effect may interact with accentuation. If we either find a main effect of syllable position on our measurements or an interaction with accentuation, we would probably proceed and refute the null hypothesis (there is no difference between stressed and unstressed syllables). This rationale can be applied to external covariates, too. For example, in prosodic studies, researchers commonly add the sex of their speakers to their models, either as a main effect or an interaction term, so as to control for the large sex-specific f_0 variability. Even though this reasoning is justified by independent factors, it represents a set of researcher degrees of freedom.

To summarize, the multidimensional nature of speech offers a myriad of different ways to look at our data. It allows us to choose dependent variables from a large pool of candidates; it allows us to measure the same dependent variable in alternative ways; and it allows us to preprocess our data in different ways by for example normalization or smoothing algorithms. Moreover, the complex interplay of different levels of speech phenomena introduced the possibility to correct or discard data during data annotation in a non-blinded fashion; it allows us to measure other variables that can be used as

covariates, mediators, or moderators. These variables could also enable further exclusion of participants (see **Figure 2** for a schematic example).

While there are often good reasons to go down one forking path rather than another, the sheer amount of possible ways to analyze speech comes with the danger of exploring many of these paths and picking those that yield a statistically significant result. Such practices, exactly because they are often unintentional, are problematic because they drastically increase the rate of false positives.

It is important to note that these researcher degrees of freedom are not restricted to speech production studies. Within more psycholinguistically oriented studies in our field, we can for example examine online speech perception by monitoring eye movements (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) or hand movements (Spivey, Grosjean, & Knoblich, 2005). There are many different informative relationships between these measured motoric patterns and aspects of the speech signal.

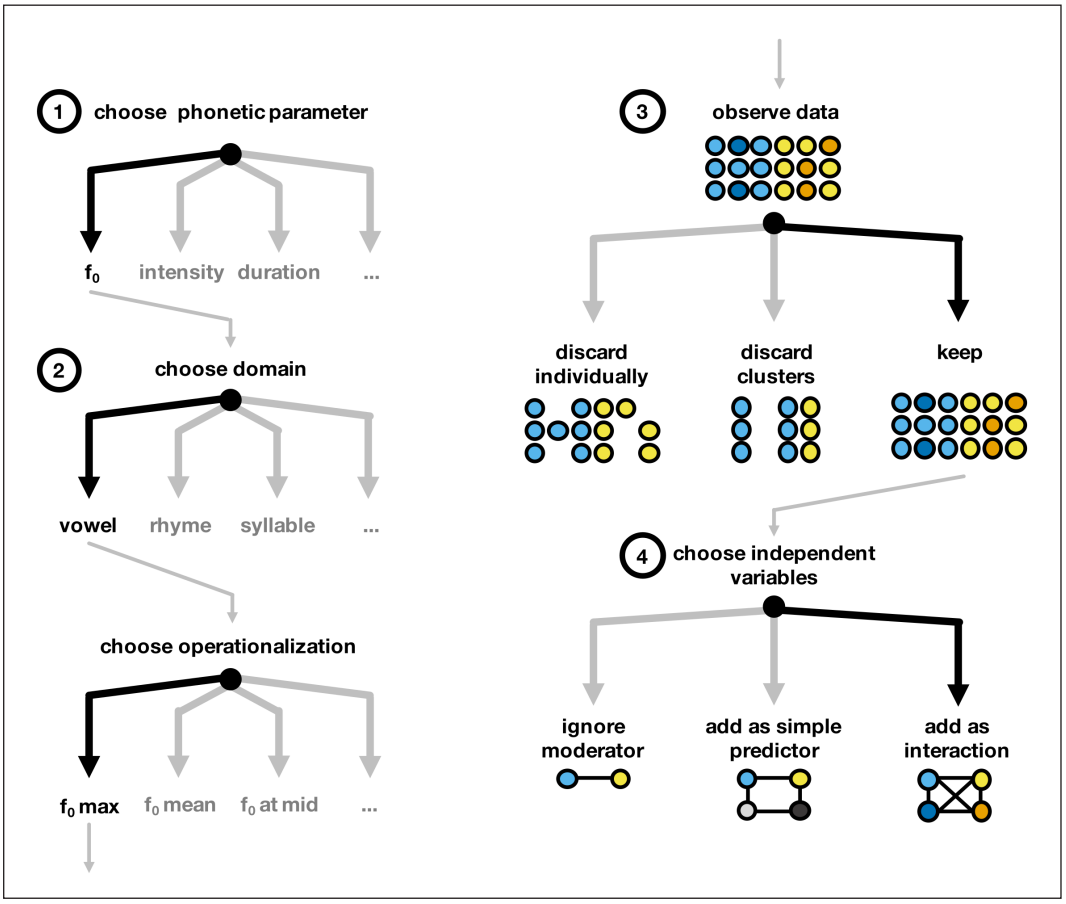

Figure 2: Schematic example of forking analytical paths as described in Section 3. First, the researchers choose from a number of possible phonetic parameters (1), then they choose in what temporal domain they measure this parameter and how they operationalize it (2). After extracting the data, the researchers must decide how to deal with aspects of speech that are orthogonal to what they are primarily interested in (say the difference between stressed and unstressed syllables: blue vs. yellow circles, [3]). For example, speakers might produce words either with or without a pitch accent (light vs. dark circles). They can discard undesired observations (e.g., discard all deaccented target words), discard clusters (e.g., discard all contrasting pairs that contain at least one deaccented target word), or keep the entire data set. If they decide to keep them, the question arises as whether to include these moderators or not and if so whether to include them as a simple predictor or add them to an interaction term (4). Note that these analytical decisions are just a subset of all possible researcher degrees of freedom at each stage.

For example, in eye tracking studies we can investigate different aspects of the visual field (foveal, parafoveal, peripheral), we can look at the number and time course of different fixations of a region (e.g., first, second, third), and the duration of a fixation or the sum of fixations before exiting/entering a particular region (see von der Malsburg & Angele, 2017, for a discussion).

More generally, the issue of researcher degrees of freedom is relevant for all quantitative scientific disciplines (see Wicherts et al., 2016). As of yet, we have not discussed much of the analytical flexibility that is specific to statistical modeling: There are choices regarding the model type, model architecture, and the model selection procedure, all of which come with their own set of researcher degrees of freedom. Having these choices is not problematic *per se*, but unintentionally exploring these choices before a final analytical commitment has been made may inflate false positive rates. Depending on the outcome and our preconception of what we expect to find, confirmation and hindsight bias may lead us to believe there are justified ways to look at the data in one particular way, until we reach a satisfying (significant) result. Again, this is a general issue in scientific practices and applies to other disciplines, too. However, given the multidimensionality of speech as well as the intricate interaction of different levels of speech behavior, the possible unintentional exploitation of researcher degrees of freedom is particularly ubiquitous in phonetic research. To demonstrate the possible severity of the issue, the following section presents a simulation that estimates false positive rates based on analytical decisions that speech production studies commonly face.

4. Simulating researcher degrees of freedom exploitation

In this section, a simulation is presented which shows that exploiting researcher degrees of freedom increases the probability of false positives, i.e., erroneously rejecting the null hypothesis. The simulation was conducted in R (R Core Team, 2016)[4] and demonstrates the effect of two different sets of researcher degrees of freedom: testing multiple dependent variables and adding a binomial covariate to the model (for similar simulations, see e.g., Barr, Levy, Scheepers, & Tily, 2013; Winter, 2011, 2015). Additionally, the correlation between dependent variables is varied with one set of simulations assuming three entirely independent measurements ($r = 0$), and one set of simulations assuming measurements that are highly correlated with each other ($r = 0.5$). The script to reproduce these simulations is publicly available here (http://osf.io/6nsfk).[5]

The simulation is based on the following hypothetical experiment: A group of researchers analyzes a speech production data set to test the hypothesis that stressed and unstressed syllables are phonetically different from each other. They collect data from 64 speakers producing words with either stress category and they measure one, two, or three acoustic parameters (e.g., vowel duration, intensity, f_0). Speakers vary regarding the intonational form of their utterances with approximately half of the speakers producing a pitch accent on the target word and the other half deaccenting the target word.

As opposed to a real-world scenario, we know the true underlying effect, since we draw values from a normal distribution around a mean value that we specify. In the present simulation, there is no difference between stressed and unstressed syllables in the 'population,' i.e., values for stressed and unstressed syllables are drawn from the same underlying distribution. However, due to random sampling (i.e., randomly picking a subset of values from the entirety of values), there are always going to be small differences

[4] The script utilizes the MASS package (Venables & Ripley, 2002) and the tidyverse library (Wickham, 2017).
[5] The simulation was inspired by Simmons et al.'s (2011) simulation which is publicly available here: https://osf.io/a67ft/.

between stressed and unstressed syllables in any given sample. Whether the word carried an accent or not is also randomly assigned to data points. In other words, there is no true effect of stress, neither is there an effect of accent on the observed productions.

We simulated 10,000 data sets and tested for the effect of stress on dependent variables under different scenarios. In our hypothetical scenarios, the following researcher degrees of freedom were explored:

(i) Instead of measuring a single dependent variable to refute the null hypothesis, the researchers measured three dependent variables. Measuring multiple aspects of the signal is common practice within the phonetic literature. In fact, measuring only three dependent variables is probably on the lower end of what researchers commonly do. However, phonetic aspects are often correlated, some of them potentially being generated by the same biomechanical mechanisms. To account for this aspect of speech, these variables were generated as either being entirely uncorrelated ($r = 0$) or being highly correlated with each other ($r = 0.5$).

(ii) The researchers explored the effect of the accent covariate and tested whether inclusion of this variable as a main effect or as an interacting term with stress yields a significant effect (or interaction effect). Accent (0,1) was randomly assigned to data rows with a probability of 0.5, that is, there is no inherent relationship with the measured parameters.

These two hypothetical scenarios correspond to researcher degrees of freedoms, discussed in Section 3.1 and Section 3.4 (see **Figures 1** and **2**). Based on these scenarios, the researchers ran simple linear models on respective dependent variables with stress as a predictor (and accent as well as their interaction in scenario ii). The simulation counts the number of times the researchers would obtain at least one significant result (p-value < 0.05) exploring the garden of forking paths described above.

The simulation (as well as the scenario it is based on) is admittedly a simplification of real-world scenarios and might therefore not be representative. As discussed above, the number of researcher degrees of freedom are manifold in phonetic investigations, therefore any simulation has to be a simplification of the real state of affairs. For example, the simulation is based on a regression model for a between-subject design, therefore not taking within-subject variation into account. It is thus important to note that the 'true' false positive rates in any given experimental scenario are not expected to be precisely as reported, or that the relative effects of the various researcher degrees of freedom are similar to those found in the simulation. Despite these necessary simplifications, the generated numbers can be considered informative and are intended to illustrate the underlying principles of how exploiting researcher degrees of freedom can impact the false positive rate.

We should expect about 5% significant results for an alpha level of 0.05 (the commonly accepted threshold in NHST). Knowing that there is no stress difference in our simulated population, any significant result is by definition a false positive. Given the proposed alpha level, we expect that—on average—1 out of 20 studies will show a false positive. **Figure 3** illustrates the results based on 10,000 simulations.

The baseline scenario (only one measurement, no covariate added) provides the expected base rate of false positives (5%, see row 1 in **Figure 3**). Departing from this baseline, the false positive rate increases substantially (rows 2–6). Looking at the results for the uncorrelated measurements first (light/red bars), results of the simulation indicate that as the researchers measure three dependent variables (rows 2–3), e.g., three possible acoustic exponents, they obtain a significant effect in 14.6% of all cases. In other words, by testing

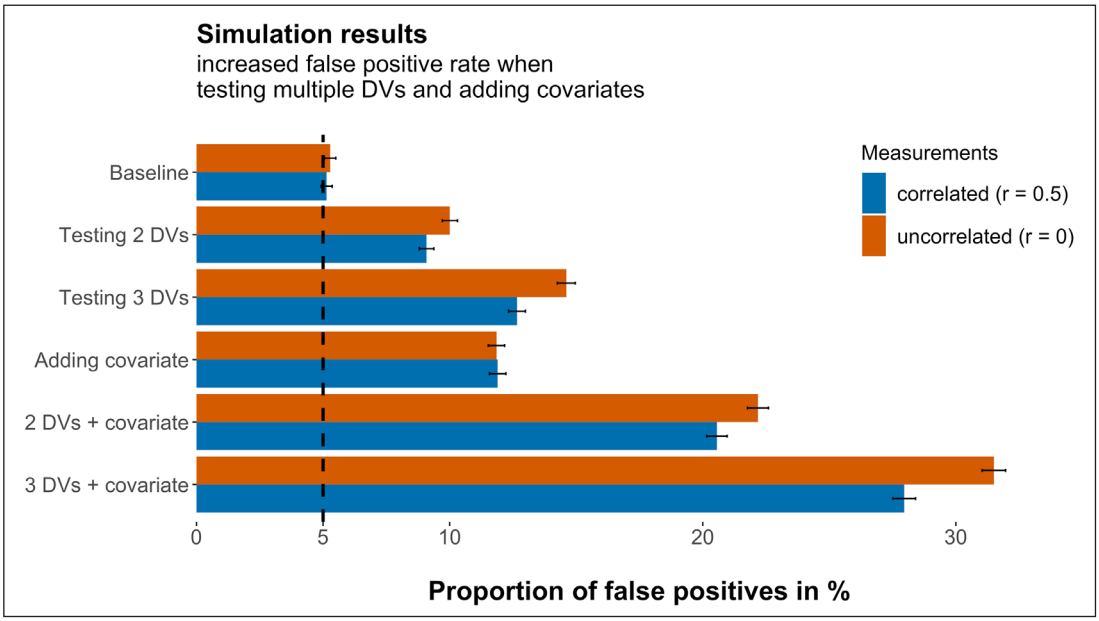

Figure 3: The x-axis depicts the proportion of simulations for which at least one attempted analysis was significant at a 0.05 level. Error bars correspond to standard deviations. Light/red bars indicate the results for uncorrelated measures, dark/blue bars indicate results for highly correlated measures (r = 0.5). Results were obtained for running three separate linear models for each dependent variable (row 1–3); for running one linear model with the main effect only, one model allowing for covariance with an accent main effect, and one model with an accent interaction (row 4, a significant effect is reported if the effect of the condition or its interaction with accent was significant in any of these analyses). Row 5–6 combine multiple dependent variables with adding a covariate. The dashed line indicates the expected false positive base rate of 5%.

three different measurements, the false positive rate increases by a factor of almost 3. When researchers explore the possibility of the effect under scrutiny covarying with a random binomial covariate, they obtain a significant effect in 11.9% of cases, increasing the false positive rate by a factor of 2.3. These numbers may strike one as rather low but they are based on only a small number of researcher degrees of freedom. In real-world scenarios, some of these choices are not entirely random but potentially informed by the researcher's preconceptions and expectations, potentially clouded by cognitive biases.

Moreover, often we unintentionally exploit multiple different researcher degrees of freedom at the same time. If the researchers combine measuring three different acoustic exponents *and* run additional analyses for the added covariate, the false error rate increases to 31.5%, i.e., the false positive rate is more than six times larger than our agreed rate of 5%.

One might argue that the acoustic measurements we take are highly correlated, making the choice between them not as impactful since they 'measure the same thing.' To explore the impact of correlated measurements, the same simulations were run with highly correlated measurements (r = 0.5, see dark/blue bars in **Figure 3**). Although the false positive rate is slightly lower than for the uncorrelated measures, we still obtain a large number of false positives. If the researchers measure three different acoustic exponents and run additional analyses for the added covariate, the false error rate is 28%. Thus, despite the measurements being highly correlated, the false error rate is still very high (see also von der Malsburg & Angele, 2017 for a discussion of eye-tracking data). In order to put these numbers into context, imagine a journal that publishes 10 papers per issue

and each paper reports only one result. In the worst-case scenario, i.e., a scenario in which the null hypothesis is true and researcher degrees of freedom are explored as described above, three to four papers of this issue would report a false positive.

Keep in mind that the presented simulation is necessarily a simplification of real-world data sets. The true false positive rates might differ from those presented here. However, even within a simplified scenario, the present simulation serves as a proof of concept that exploiting researcher degrees of freedom can impact the false positive rate quite substantially.

Another limitation of the present simulation (and of most simulations of this nature) is that it is blind to the direction of the effect. Here, we counted every *p*-value below our set alpha level as a false positive. In real-world scenarios, however, one could argue that we often have directional hypotheses. For example, we expect unstressed syllables to exhibit 'significantly' shorter durations/lower intensity/lower f_0, etc. It could be argued that finding a significant effect going in the 'wrong' direction is potentially less likely to be considered trustworthy evidence (i.e., unstressed syllables being longer). However, due to cognitive biases such as hindsight bias (Fischhoff, 1975) and overconfident beliefs in the replicability of significant results (Vasishth, Mertzen, Jäger, & Gelman, 2018) the subjective expectedness of effect directionality might be misleading.

The issues discussed above are general issues of data analysis operating in a particular inferential framework and they are not specific to phonetic sciences. However, we need to be aware of these issues and we need to have an open discourse about them. It turns out that one of the aspects that makes our scientific object so interesting—the complexity of speech—can pose a methodological burden.

5. Possible remedies

The false positive rates demonstrated above are not an inevitable fate. In the following, I will discuss possible strategies to elude the threat posed by researcher degrees of freedom, focusing on five different topics: I discuss the possibility of adjusting the alpha level as a way to reduce false positives (Section 5.1). Alternatively, I propose that drawing a clear line between exploratory and confirmatory analyses (Section 5.2) and committing to analytical decisions prior to data collection with preregistered protocols (Section 5.3) can limit the number of false positives. Additionally, a valuable complementary practice may lie in open, honest, and transparent reporting on how and what was done during the data analysis procedure (Section 5.4). While transparency cannot *per se* limit the exploitation of researcher degrees of freedom, it can facilitate their detection. Finally, it is argued that direct replications are our strongest strategy against false positives and researcher degrees of freedom (Section 5.5).

5.1. Adjusting the significance threshold

The increased false positive rates as discussed in this paper are closely linked to null hypothesis significance testing which, in its contemporary form, constitutes a dichotomous statistical decision procedure based on a preset threshold (the alpha level). At first glance, an obvious solution to reduce the likelihood of obtaining false positives is to adjust the alpha level, either by correcting it for relevant researcher degrees of freedom or by lowering the decision threshold for significance *a priori*.

First, one could correct the alpha level as a function of the number of exploited researcher degrees of freedom. This solution has mainly been discussed in the context of multiple comparisons, in which the researcher corrects the alpha level threshold according to the number of tests performed (Benjamini & Hochberg, 1995; Tukey, 1953). If we were to measure three acoustic parameters to test a global null hypothesis, which can be refuted by

a single statistically significant result, we would lower the alpha level to account for three tests. These corrections can be done for example via the Bonferroni or the Šidák method.[6]

One may object that corrections for multiple testing are not reasonable in the case of speech production on the grounds that acoustic measures are usually correlated. In this case, correcting for multiple tests may be too conservative. However, as demonstrated in Section 4 and discussed by von der Malsburg and Angele (2017), multiple testing of highly correlated measures leads to false positive rates that are nearly as high as for independent multiple tests. Thus, a multiple comparisons correction is necessary even with correlated measures in order to obtain the conventional false positive rate of 5%. However, given the large analytical decision space we have discussed above, it remains unclear as to how much to correct the alpha level for other individual researcher degree of freedom. Moreover, given that speech production experiments usually yield a limited amount of data, strong alpha level corrections can drastically inflate false negatives (Type II errors, e.g., Thomas et al., 1985), i.e., erroneously failing to reject the null hypothesis.

Complementarily, one could pose a more conservative alpha level *a priori*. Benjamin et al. (2018) recently made such a proposal, recommending to lower our commonly agreed alpha level from $p \leq 0.05$ to $p \leq 0.005$. All else being equal, lowering the alpha level will reduce the absolute number of false positives (e.g., in the simulation above, we would only obtain around 0.5% to 3% of false positives across scenarios). While some researchers articulated their concerns that lowering the commonly accepted threshold for significance comes with important drawbacks such as an increase in false negatives and increased resource costs (Amrhein & Greenland, 2018; Lakens et al., 2018, but see de Ruiter, 2018), it can certainly help us reduce false positives by raising the bar as to what counts as significant and what does not.

In sum, correcting the alpha level or setting a lower alpha level threshold for significance can be helpful strategies to control for false positive rates in a conservative way. As with every practice, alpha level adjustments have their own drawbacks. It sometimes remains unclear as how to exactly adjust the alpha level in a non-conservative way. Moreover, alpha level adjustment can increase false negative rates.

5.2. Flagging analyses as exploratory vs. confirmatory

One important remedy to the issue of researcher degrees of freedom is to draw a clear line between exploratory and confirmatory analyses, two conceptually separate phases in scientific discovery (Box, 1976; Tukey, 1980; de Groot, 2014; see Nicenboim, Vasishth, Engelmann, & Suckow, 2018b, and Roettger, Winter, & Baayen, accepted, for recent discussions related to linguistic research). In an exploratory analysis, we observe patterns and relationships which lead to the generation of concrete hypotheses as to how these observations can be explained. These hypotheses can then be challenged by collecting new data (e.g., in controlled experiments). Putting our predictions under targeted scrutiny helps us revise our theories based on confirmatory analyses. Our revised models can then be further informed by additional exploration of the available data. This iterative process of alternating exploration and confirmation advances our knowledge. This is hardly news to the reader. However, quantitative research in science in general, as well as in our field in particular, often blurs the line between these two types of data analysis.

Exploratory and confirmatory analyses should be considered complementary to each other. Unfortunately, when it comes to publishing our work, they are not weighted equally.

[6] One approach is to use an additive (Bonferroni) inequality: For n tests, the alpha level for each test is given by the overall alpha level divided by n. A second approach is to use a multiplicative inequality (Šidák): For n tests, the alpha level for each test is calculated by taking 1 minus the n^{th} root of the complement of the overall alpha level.

Confirmatory analyses have a superior status, determining the way we frame our papers and the way funding agencies demand successful proposals to look like. This asymmetry can have harmful consequences which I have discussed already: *HARKing* and *p-hacking*. It may also incentivize researchers to sidestep clear-cut distinctions between exploratory and confirmatory findings. The publication apparatus forces us into a confirmatory mindset, while we often want to explore the data and generate hypotheses. For example, we want to explore what the most important phonetic exponents of a particular functional contrast are. We may not necessarily have a concrete prediction we want to test at this stage, but we want to understand patterns in speech with respect to their function. Exploratory analyses are necessary to establish standards as to how aspects of speech relate to linguistic, cognitive, and social variables. Once we have established such standards, we can agree to only look at relevant phonetic dimensions, reducing the analytical flexibility with regard to what and how to measure (Sections 3.1–3.2).

Researcher degrees of freedom mainly affect the confirmatory part of scientific discovery; they do not restrict our attempts to explore our data. But claims based on exploration should be cautious. After having looked at 20 acoustic dimensions, any seemingly systematic pattern may be spurious. Instead, this exploratory step should generate new hypotheses which we then can confirm or disconfirm using a new data set. In many experiments, prior to data collection, it may not be clear how a functional contrast may phonetically manifest itself. Presenting such exploratory analyses as confirmatory may hinder replicability and may give a false feeling of certainty regarding the results (Vasishth et al., 2018). In a multivariate setting, which is the standard setting for phonetic research, there are multiple dimensions to the data that can inform our theories. Exploring these dimensions may often be more valuable than just a single confirmatory test of a single hypothesis (Baayen, Vasishth, Kliegl, & Bates, 2017). These cases make the distinction between confirmatory and exploratory analyses so important. We should explore our data. Yes. Yet we should not pretend we are testing concrete hypotheses when doing so.

Although our academic incentive system makes drawing this line difficult, journals have started to become aware of this issue and have started to create incentives to explicitly publish exploratory analyses (for example Cortex, see McIntosh, 2017). One way of ensuring a clear separation between exploratory and confirmatory analyses are preregistrations and registered reports (Nosek, Ebersole, DeHaven, & Mellor, 2018; Nosek & Lakens, 2014).

5.3. Preregistrations and registered reports

A preregistration is a time-stamped document in which researchers specify exactly how they plan to collect their data and how they plan to conduct their confirmatory analyses. Such reports can differ with regard to the details provided, ranging from basic descriptions of the study design to detailed procedural and statistical specifications up to the publication of scripts.

Preregistrations can be a powerful tool to reduce researcher degrees of freedom because researchers are required to commit to certain decisions prior to observing the data. Additionally, public preregistration can at least help to reduce issues related to publication bias, i.e., the tendency to publish positive results more often than null results (Franco, Malhotra, & Simonovits, 2014; Sterling, 1959), as the number of failed attempts to reject a hypothesis can be tracked transparently (if the studies were conducted).

There are several websites that offer services and/or incentives to preregister studies prior to data collection, such as AsPredicted (AsPredicted.org) and the Open Science Framework (osf.io). These platforms allow us to time-log reports and either make them publicly available or grant anonymous access only to a specific group of people (such as reviewers and editors during the peer-review process).

A particular useful type of preregistration is a peer-reviewed registered report, which an increasing number of scientific journals have adopted already (Nosek et al., 2018; Nosek & Lakens, 2014; see cos.io/rr for a list of journals that have adopted this model).[7] These protocols include the theoretical rationale of the study and a detailed methodological description. In other words, a registered report is a full-fledged manuscript minus the result and discussion section. These reports are then critically assessed by peer reviewers, allowing the authors to refine their methodological design. Upon acceptance, the publication of the study results is in-principle guaranteed, no matter whether the results turn out to provide evidence for or against the researcher's predictions.

For experimental phonetics, a preregistration or registered report would ideally include a detailed description of what is measured and how exactly it is measured/operationalized, as well as a detailed catalogue of objective inclusion criteria (in addition to other key aspects of the method including all relevant researcher degrees of freedom related to preprocessing, postprocessing, statistical modelling, etc.; see Wicherts et al., 2016). Committing to these decisions prior to data collection can reduce the danger of unintentionally exploiting researcher degrees of freedom.

At first sight, there appear to be several challenges that come with preregistrations (see Nosek et al., 2018, for an exhaustive discussion). For example, after starting to collect data, we might realize that our preset exclusion criteria do not capture an important behavioral aspect of our experiment (e.g., some speakers may produce undesired phrase-level prosodic patterns which we did not anticipate). These patterns interfere with our research question. Deviations from our data collection and analysis plan are common. In this scenario, we could change our preregistration and document these changes alongside our reasons as to why and when we have made these changes (i.e., after how many observations). This procedure still provides substantially lower risk of cognitive biases impacting our conclusions compared to a situation in which we did not preregister at all.

Researchers working with corpora may object that preregistrations cannot be applied to their investigations because their primary data have already been collected. But preregistration of analyses can still be performed. Although, ideally, we limit researcher degrees of freedom prior to having seen the data, we can (and should) preregister analyses after having seen pilot data, parts of the study, or even whole corpora. When researchers generate a hypothesis that they want to confirm with a corpus data set, they can preregister analytic plans and commit to how evidence will be interpreted before analyzing the data.

Another important challenge when preregistering our studies is predicting appropriate inferential models. Preregistering a data analysis necessitates knowledge about the nature of the data. For example, we might preregister an analysis assuming that our measurements are generated by a Gaussian process. After collecting our data, we might realize the data have heavy right tales, calling for a log-transformation; thus, our preregistered analysis might not be appropriate. One solution to this challenge is to define data analytical procedures in advance that allow us to evaluate distributional aspects of the data and potential data transformations irrespective of the research question. Alternatively, we could preregister a decision tree. This may actually be tremendously useful for people using hierarchical linear models. When using appropriate random effect structures (see Barr et al., 2013; Bates et al., 2015), these models are known to run into convergence issues (e.g., Kimball, Shantz, Eager, & Roy, 2018). To remedy such convergence issues, a common strategy is to drop complex random effect terms incrementally (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Since we do not know whether a model will converge or not in advance, a concrete plan of how we reduce model complexity can be preregistered in advance.

[7] Note that there are only a few journals on quantitative linguistics that have adopted registered reports.

Preregistrations and registered reports help us draw a line between the hypotheses we intend to test and data exploration (see **Figure 4**). Any exploration of the data beyond the preregistered analysis have to be considered as hypotheses-generating only. If we are open and transparent about this distinction and ideally publicly indicate where we draw the line between the two, we can limit false positives due to researcher degrees of freedom exploitation in our confirmatory analyses and commit more honestly to subsequent exploratory analyses.

5.4. Transparency

The credibility of scientific findings is mainly rooted in the evidence supporting it. We assess the validity of this evidence by constantly reviewing and revising our methodologies, and by extending and replicating findings. This becomes difficult if parts of the process are not transparent or cannot be evaluated. For example, it is difficult to evaluate whether exploitation of researcher degrees of freedom is an issue for any given study if the authors are not transparent about when they made which analytical decisions. We end up having to *trust* the authors. Trust is good, control is better. We should be aware and open about researcher degrees of freedom and communicate this aspect of our data analyses to our peers as honestly as we can. For example, if we have measured and analyzed ten phonetic parameters to establish whether the final syllable of a word is more prominent than the prefinal syllable, we should provide this information: If we have run analyses with and without gender as a covariate, we should say so and discuss the results of these analyses. An open, honest, and transparent research culture is desirable. As argued above, public preregistration can facilitate transparency of what analytical decisions we made and when we have made them. Being transparent does not, of course, prevent *p*-hacking, HARKing or other exploitations of researcher degrees of freedom, but it makes these harmful practices *detectable*.

For our field, transparency with regard to our analysis has many advantages (e.g., Nicenboim et al., 2018b).[8] As analysis is subjective in the sense that it incorporates the researcher's beliefs and assumptions about a study system (McElreath, 2016) the only way to make analyses objectively assessable is to be transparent about this aspect of empirical work. Transparency then allows other researchers to draw their own conclusions as to which researcher degrees of freedom were present and how they may have affected the original conclusion.

In order to facilitate such transparency, we need to agree on how to report aspects of our analyses. While preregistrations and registered reports lead to better discoverability of researcher degrees of freedom, they do not necessarily allow us to systematically evaluate them. We need institutionalized standards, as many other disciplines have already developed. There are many reporting guidelines that offer standards for reporting methodological choices (see the Equator Network for an aggregation of these guidelines: http://www.equator-network.org/). Systematic reviews and meta analyses such as Gordon and Roettger (2017) and Roettger and Gordon (2017) can be helpful departure points to create an overview of possible analytical decisions and their associated degrees of freedom (e.g., what is measured; how it is measured, operationalized, processed, and extracted; what data are excluded; when and how are they excluded, etc.). Such guidelines, however,

[8] Beyond sharing data tables and analysis scripts, it would be desirable to share raw acoustic or articulatory files. However, making these types of data available relies on getting permission from participants in advance (as acoustic data are inherently identifiable). Making raw speech production data available to the community would greatly benefit evidence accumulation in our field. We can share these data on online repositories such as, for example, OSCAAR (the Online Speech/Corpora Archive and Analysis Resource: https://oscaar.ci.northwestern.edu/).

are only effective when a community agrees on their value and applies them including journals, editors, reviewers, and authors.

5.5. Direct replications

The above discussed remedies help us to either limit the exploitation of researcher degrees of freedom or make them more detectable. However, none of these strategies is a fool-proof protection against false positives. To ultimately avoid the impact of false positives on the scientific record, we should increase our efforts to directly replicate previous research, defined here as the repetition of the experimental methods that led to a reported finding.

The call for more replication is not original. Replication has always been considered a tremendously important aspect of the scientific method (e.g., Campbell, 1969; Kuhn, 1962; Popper, 1934/1992; Rosenthal, 1991) and in recent coordinated efforts to replicate published results, the social sciences uncovered unexpectedly low replicability rates, a state of affairs that has been coined the 'replication crisis.' For example, the Open Science Collaboration (2015) tried to replicate 100 studies that were published in three high-ranking psychology journals. They assessed whether the replications and the original experiments yielded the same result and found that only about one third to one half of the original findings (depending on the definition of replication) were also observed in the replication study. This lack of replicability is not restricted to psychology. Concerns about the replicability of findings have been raised for medical sciences (e.g., Ioannidis, 2005), neuroscience (Wager, Lindquist, Nichols, Kober, & van Snellenberg, 2009), genetics (Hewitt, 2012), cancer research (Errington et al., 2014), and economics (Camerer et al., 2016).

Most importantly, it is a very real problem for quantitative linguistics, too. For example, Nieuwland et al. (2018) recently tried to replicate a seminal study by DeLong, Urbach, and Kutas (2005) which is considered a landmark study for the predictive processing literature and which has been cited over 500 times. In their preregistered multi-site replication attempt (9 laboratories, 334 subjects), Nieuwland et al. were not able to replicate some of the key findings of the original study.

Stack, James, and Watson (2018) recently failed to replicate a well-cited study on rapid syntactic adaptation by Fine, Jaeger, Farmer, and Qian (2013). After failing to find the original effect in an extension, they went back and directly replicated the original study with appropriate statistical power. They found no evidence for the original effect.

Possible reasons for the above cited failures to replicate are manifold. As has been argued here, exploitation of researcher degrees of freedom is one of the reasons why there is a large number of false positives. Combined with other statistical issues such as low power (e.g., for recent discussion see Kirby & Sonderegger, 2018; Nicenboim et al., 2018a), violation of the independence assumption (Nicenboim & Vasishth, 2016; Winter, 2011, 2015), and the 'significance' filter (i.e., treating results publishable because $p < 0.05$ leads to overoptimistic expectations of replicability; see Vasishth et al., 2018), it is to be expected that there are a large number of experimental phonetic findings that may not stand the test of time.

The above replication failures sparked a tremendously productive discourse throughout the quantitative sciences and led to quick methodological advancements and best practice recommendations. For example, there are several coordinated efforts to directly replicate important findings by multi-site projects such as the ManyBabies project (Frank et al., 2017) and Registered Replication Reports (Simons, Holcombe, &, Spellman, 2014). These coordinated efforts can help us put theoretical foundations on a firmer footing. However,

the logistic and monetary resources associated with such large-scale projects are not always pragmatically feasible for everyone in the field.

Replication studies are not very popular because the necessary time and resource investment are not appropriately rewarded in contemporary academic incentive systems (Koole & Lakens, 2012; Makel, Plucker, & Hegarty, 2012; Nosek, Spies, & Motyl, 2012). Both successful replications (Madden, Easley, & Dunn, 1995) and repeated failures to replicate (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012) are rarely published, and if they are published they are usually published in less prestigious outlets than the original findings. To overcome the asymmetry between the cost of direct replication studies and the presently low academic payoff for it, we as a research community must re-evaluate the value of direct replications. Funding agencies, journals, editors, and reviewers should start valuing direct replication attempts, be it successful replications or replication failures, as much as they value novel findings. For example, we could either dedicate existing journal space to direct replications (e.g., as an article type) or create new journals that are specifically dedicated to replication studies. For example, the Royal Society Open Science has recently initiated an interesting new publication model. For their Psychology and Cognitive Neuroscience section, they guarantee to publish any close replication of any article published in their own and other journals (https://blogs.royalsociety.org/publishing/reproducibility-meets-accountability/). Thus, if the journal agrees to publish a study, it becomes responsible for publishing direct replications of that study, too.

As soon as we make publishing replications easier, more researchers will be compelled to replicate both their own work and the work of others. Only by replicating empirical results and evaluating the accumulated evidence can we substantiate previous findings and extend their external validity.

6. Summary and concluding remarks

This article has discussed researcher degrees of freedom in the context of quantitative phonetics. Researcher degrees of freedom concern all possible analytical choices that may influence the outcome of our analysis. In a null-hypothesis-significance testing framework of inference, intentional or unintentional exploitation of researcher degrees of freedom can have a dramatic impact on our results and interpretations, increasing the likelihood of obtaining false positives. Quantitative phonetics faces a large number of researcher degrees of freedom due to its scientific object being inherently multidimensional and exhibiting complex interactions between many covarying layers of speech. A Type-I error simulation demonstrated substantial false error rates when combining just two researcher degrees of freedom such as testing more than one phonetic measurement, and including a speech-relevant covariate in the analysis. It has been argued that combined with common cognitive fallacies, unintentional exploitation of researcher degrees of freedom introduces strong bias and poses a serious challenge to quantitative phonetics as an empirical science.

Several potential remedies for this problem have been discussed (see **Figure 4**). When operating in the NHST statistical framework, we can reconsider our preset threshold for significance. We should draw an explicit line between confirmatory and exploratory analyses. One way to enforce such a clear line are preregistrations or registered reports, i.e., records of the experimental design and the analysis plan that are committed prior to data collection and analysis. While preregistration offers better detectability of researcher degrees of freedom, standardized reporting guidelines and transparent reporting might facilitate a more objective assessment of these researcher degrees of freedom by other

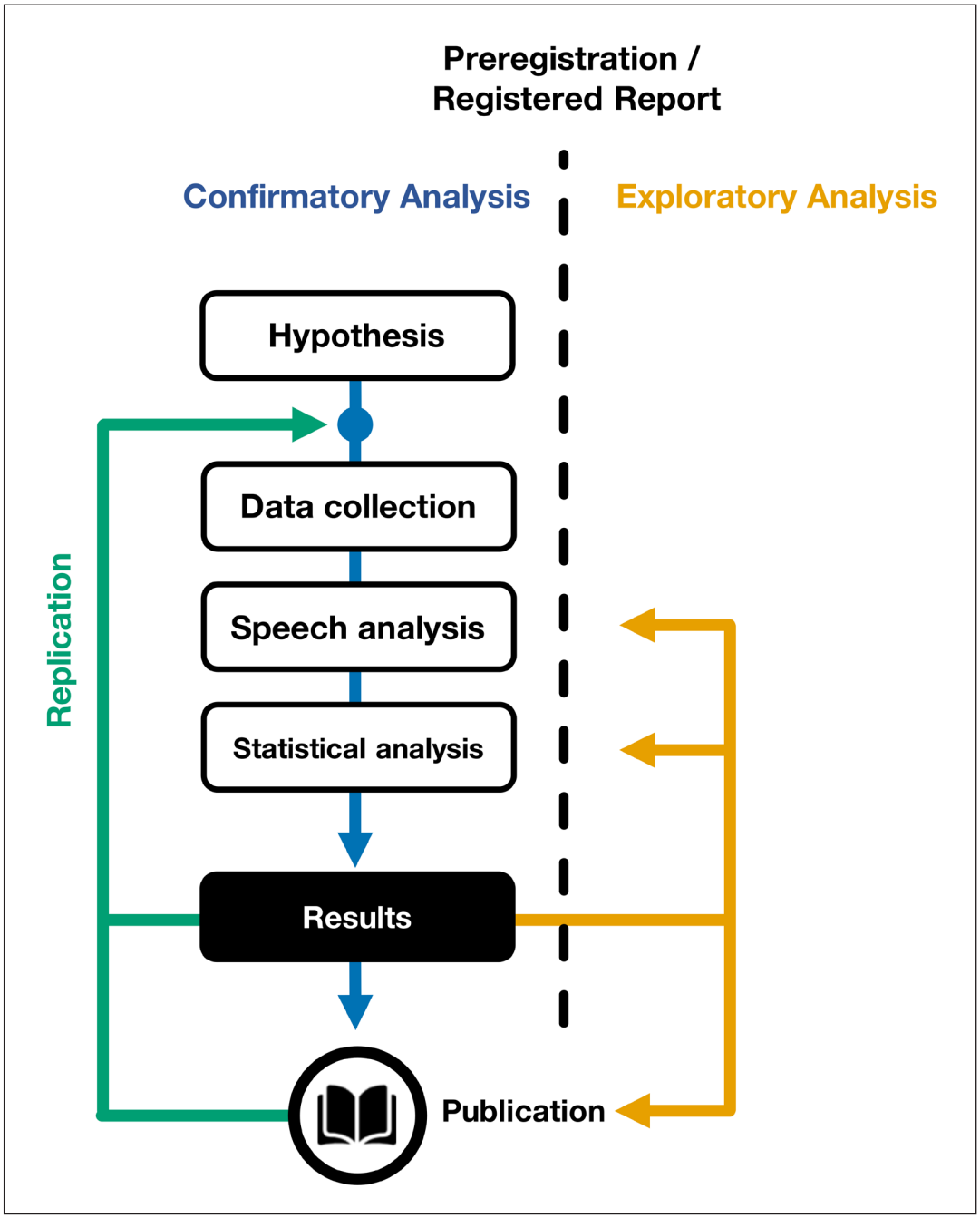

Figure 4: Schematic depiction of the decision procedure during data analysis that limits false positives: Prior to data collection, the researcher commits to an analysis pipeline via preregistration/registered reports, leading to a clear separation of confirmatory (blue arrows) and exploratory analysis (yellow arrows). The analysis is executed accordingly and the results are interpreted with regard to the confirmatory analysis. After the confirmatory analysis, the researcher can revisit the decision procedure and explore the data. The interpretation of the confirmatory analysis and potential insights gained from the exploratory analysis are published alongside an open and transparent track record of all analytical steps (preregistration, code, and data for both confirmatory and exploratory analyses). Finally, either prior to publication or afterwards, the study is directly replicated (green arrow) by either the same research group or independent researchers in order to substantiate the results.

researchers. Yet all of these proposals come with their own limitations and challenges. A complementary strategy to limit false positives lies in direct replications, a form of research that is unfortunately not well rewarded within the present academic system.

As a community, we need to openly discuss such issues and find feasible solutions to them. Possible solutions must not only be practical from a logistic perspective but should also avoid punishing rigorous methodology within our academic incentive system. Explicitly labeling our work as exploratory, being transparent about potential bias due to researcher degrees of freedom, or running direct replications may make it more difficult to be rewarded for our work (i.e., by being able to publish our studies in prestigious journals). Thus, authors, reviewers, and editors alike need to be aware of these methodological challenges. The present paper was conceived in the spirit of such an open discourse. Thus, the single most powerful solution to methodological challenges as described in this paper is engaging in a critical and open discourse about our methods and analyses.

Acknowledgements

I would like to thank Aviad Albert, Eleanor Chodroff, Emily Cibelli, Jennifer Cole, Tommy Denby, Matt Goldrick, James Kirby, Nicole Mirea, Bruno Nicenboim, Limor Raviv, Shayne Sloggett, Lukas Soenning, Mathias Stoeber, Bodo Winter, two anonymous reviewers, and the handling editor for valuable feedback on earlier drafts of this paper. All remaining errors are my own.

Competing Interests

The author has no competing interests to declare.

References

Amrhein, V., & Greenland, S. 2018. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, *2*(1), 4. DOI: https://doi.org/10.1038/s41562-017-0224-0

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. DOI: https://doi.org/10.1016/j.jml.2016.11.006

Bakker, M., & Wicherts, J. M. 2011. The (mis) reporting of statistical results in psychology journals. *Behavior research methods*, *43*(3), 666–678. DOI: https://doi.org/10.3758/s13428-011-0089-5

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. 2015. Parsimonious mixed models. *arXiv Preprint*. arXiv:1506.04967.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., & Cesarini, D. 2018. Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. DOI: https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Benkí, J. R. 2001. Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, *29*(1), 1–22. DOI: https://doi.org/10.1006/jpho.2000.0128

Box, G. E. P. 1976. Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. DOI: https://doi.org/10.1080/01621459.1976.10480949

Brugger, P. 2001. From haunted brain to haunted science. In: Houran, J., & Lange, R. (eds.), *Hauntings and poltergeists: Multidisciplinary perspectives*, 195–213. McFarland.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., & Heikensten, E. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. DOI: https://doi.org/10.1126/science.aaf0918

Campbell, D. T. 1969. Reforms as experiments. *American Psychologist*, *24*(4), 409–429. DOI: https://doi.org/10.1037/h0027982

Cangemi, F. 2015. *Prosodic detail in Neapolitan Italian*. Language Science Press: Berlin. DOI: https://doi.org/10.26530/oapen_533874

Cho, T., & Keating, P. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485. DOI: https://doi.org/10.1016/j.wocn.2009.08.001

Cummins, F. 2012. Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes*, *27*(10), 1525–1549. DOI: https://doi.org/10.1080/01690965.2011.615220

Dawson, E., Gilovich, T., & Regan, D. T. 2002. Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387. DOI: https://doi.org/10.1177/014616702236869

de Groot, A. D. 2014. The meaning of "significance" for different types of research [translated and annotated by E.-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, & H. L. J. van der Maas]. *Acta psychologica*, *148*, 188–194. DOI: https://doi.org/10.1016/j.actpsy.2014.02.001

DeLong, K. A., Urbach, T. P., & Kutas, M. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117–1121. DOI: https://doi.org/10.1038/nn1504

de Ruiter, J. 2018. Redefine or justify? Comments on the alpha debate. *Psychonomic bulletin & review*, 1–4. DOI: https://doi.org/10.3758/s13423-018-1523-9

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. 2012. Behavioral priming: It's all in the mind, but whose mind? *PloS one*, *7*(1), e29081. DOI: https://doi.org/10.1371/journal.pone.0029081

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. 2014. Science forum: An open investigation of the reproducibility of cancer biology research. *Elife*, *3*, e04333. DOI: https://doi.org/10.7554/eLife.04333

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. 2013. Rapid expectation adaptation during syntactic comprehension. *PloS one*, *8*(10), e77661. DOI: https://doi.org/10.1371/journal.pone.0077661

Fischhoff, B. 1975. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance*, *1*(3), 288–299. DOI: https://doi.org/10.1037/0096-1523.1.3.288

Franco, A., Malhotra, N., & Simonovits, G. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. DOI: https://doi.org/10.1126/science.1255484

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., & Lew-Williams, C. 2017. A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. DOI: https://doi.org/10.1111/infa.12182

Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. 2004. Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 58*(2), 86–95. DOI: https://doi.org/10.1037/h0085799

Gelman, A., & Carlin, J. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. DOI: https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. 2014. Ethics and statistics: The AAA tranche of subprime science. *Chance*, *27*(1), 51–56. DOI: https://doi.org/10.1080/09332480.2014.890872

Gigerenzer, G., Krauss, S., & Vitouch, O. 2004. The null ritual. In: Kaplan, D. (ed.), *The Sage handbook of quantitative methodology for the social sciences*, 391–408. Thousand Oaks, CA: Sage. DOI: https://doi.org/10.4135/9781412986311.n21

Gordon, M., & Roettger, T. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1). DOI: https://doi.org/10.1515/lingvan-2017-0007

Greenland, S. 2017. Invited commentary: The need for cognitive science in methodology. *American journal of epidemiology*, *186*(6), 639–645. DOI: https://doi.org/10.1093/aje/kwx259

Haggard, M., Ambler, S., & Callow, M. 1970. Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, *47*(2B), 613–617. DOI: https://doi.org/10.1121/1.1911936

Harrington, J., Fletcher, J., & Beckman, M. 2000. Manner and place conflicts in the articulation of accent in Australian English. In: Broe, M. (ed.), *Papers in Laboratory Phonology*, *5*, 40–55. Cambridge University Press: Cambridge.

Hastorf, A. H., & Cantril, H. 1954. They saw a game: A case study. *The Journal of Abnormal and Social Psychology*, *49*(1), 129–134. DOI: https://doi.org/10.1037/h0057880

Hawkins, S., & Nguyen, N. 2004. Influence of syllable-coda voicing on the acoustic properties of syllable-onset/l/in English. *Journal of Phonetics*, *32*(2), 199–231. DOI: https://doi.org/10.1016/s0095-4470(03)00031-7

Hewitt, J. K. 2012. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior genetics*, *42*(1), 1–2. DOI: https://doi.org/10.1007/s10519-011-9504-z

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. 1979. Phonetic explanations for the development of tones. *Language*, *55*, 37–58. DOI: https://doi.org/10.2307/412518

Ioannidis, J. P. 2005. Why most published research findings are false. *PLoS medicine*, *2*(8), e124. DOI: https://doi.org/10.1093/bioinformatics/bti536

John, L. K., Loewenstein, G., & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth tell. *Psychological Science*, *23*, 524–532. DOI: https://doi.org/10.1037/e632032012-001

Jongman, A., Wayland, R., & Wong, S. 2000. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. DOI: https://doi.org/10.1121/1.1288413

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Kimball, A. E., Shantz, K., Eager, C., & Roy, J. 2018. Beyond maximal random effects for logistic regression: Moving past convergence errors. *Journal of Quantitative Linguistics*, 1–25. DOI: https://doi.org/10.1080/09296174.2018.1499457

Kirby, J., & Sonderegger, M. 2018. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics*, *70*, 70–85. DOI: https://doi.org/10.1016/j.wocn.2018.05.005

Koole, S. L., & Lakens, D. 2012. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608–614. DOI: https://doi.org/10.1177/1745691612462586

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., & Buchanan, E. M. 2018. Justify your

alpha. *Nature Human Behaviour*, *2*(3), 168–171. DOI: https://doi.org/10.1038/s41562-018-0311-x

Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M. S., & Munhall, K. G. 2014. Movement coordination during conversation. *PLoS One*, *9*(8), e105036. DOI: https://doi.org/10.1371/journal.pone.0105036

Lindquist, E. F. 1940. *Statistical Analysis in Educational Research*. Boston: Houghton Mifflin.

Lisker, L. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, *33*(1), 42–49. DOI: https://doi.org/10.2307/410949

Lisker, L. 1986. "Voicing" in English – A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*, 3–11. DOI: https://doi.org/10.1177/002383098602900102

Lisker, L., & Abramson, A. S. 1963. Crosslanguage study of voicing in initial stops. *The Journal of the Acoustical Society of America*, *35*(11), 1889–1890. DOI: https://doi.org/10.1121/1.2142685

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. 1989. The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America*, *85*(3), 1314–1321. DOI: https://doi.org/10.1121/1.397462

Madden, C. S., Easley, R. W., & Dunn, M. G. 1995. How journal editors view replication research. *Journal of Advertising*, *24*(4), 77–87. DOI: https://doi.org/10.1080/00913367.1995.10673490

Makel, M. C., Plucker, J. A., & Hegarty, B. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. DOI: https://doi.org/10.1177/1745691612460688

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. DOI: https://doi.org/10.1016/j.jml.2017.01.001

McElreath, R. 2016. *Statistical Rethinking*: *A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Press.

McIntosh, R. D. 2017. Exploratory reports: A new article type for Cortex. *Cortex*, *96*, A1–A4. DOI: https://doi.org/10.1016/j.cortex.2017.07.014

Nicenboim, B., Roettger, T. B., & Vasishth, S. 2018a. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, *70*, 39–55. DOI: https://doi.org/10.1016/j.wocn.2018.06.001

Nicenboim, B., & Vasishth, S. 2016. Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, *10*(11), 591–613. DOI: https://doi.org/10.1111/lnc3.12207

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. 2018b. Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive science*, *42*, 1075–1100. DOI: https://doi.org/10.1111/cogs.12589

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220. DOI: https://doi.org/10.1037//1089-2680.2.2.175

Niebuhr, O., D'Imperio, M., Gili Fivela, B., & Cangemi, F. 2011. Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category. *Proceedings of the 17th International Congress of Phonetic Sciences*, 120–123. Hong Kong.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthurn, S. V. G., Bartolozzi, F., Kogan, V., Ito, A., & Mézière, D. 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. DOI: https://doi.org/10.7554/eLife.33468

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. DOI: https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. DOI: https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. DOI: https://doi.org/10.1177/1745691612459058

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, *349*(6251). DOI: https://doi.org/10.1126/science.aac4716

Popper, K. 1992. *The logic of scientific discovery*. New York: Routledge. (Original work published 1934).

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. 2015. Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, *49*, 41–54. DOI: https://doi.org/10.1016/j.wocn.2014.10.005

Raphael, L. J. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303. DOI: https://doi.org/10.1121/1.1912974

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Repp, B. H. 1979. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, *22*(2), 173–189. DOI: https://doi.org/10.1177/002383097902200207

Ritter, S., & Roettger, T. B. 2014. Speakers modulate noise-induced pitch according to intonational context. In: *Proceedings of the 7th International Conference on Speech Prosody*, 890–893. Dublin.

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J. L. 2008. The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1507–1521. DOI: https://doi.org/10.1044/1092-4388(2008/07-0173)

Roettger, T. B., & Gordon, M. 2017. Methodological issues in the study of word stress correlates. *Linguistics Vanguard*, *3*(1). DOI: https://doi.org/10.1515/lingvan-2017-0006

Roettger, T. B., Winter, B., & Baayen, H. accepted. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Accepted for publication at Journal of Phonetics*.

Rosenthal, R. 1991. Replication in behavioral research. In: Neuliep, J. W. (ed.), *Replication research in the social sciences*, 1–39. Newbury Park, CA: Sage.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366. DOI: https://doi.org/10.1037/e519702015-014

Simons, D. J., Holcombe, A. O., & Spellman, B. A. 2014. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. DOI: https://doi.org/10.1177/1745691614543974

Smaldino, P. E., & McElreath, R. 2016. The natural selection of bad science. *Royal Society Open Science*, *3*(9). DOI: https://doi.org/10.1098/rsos.160384

Spivey, M. J., Grosjean, M., & Knoblich, G. 2005. Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398. DOI: https://doi.org/10.1073/pnas.0503903102

Stack, C. M. H., James, A. N., & Watson, D. G. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & cognition*, *46*(6), 864–877. DOI: https://doi.org/10.3758/s13421-018-0808-6

Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. DOI: https://doi.org/10.1080/01621459.1959.10501497

Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, *7*, 1074–1095. DOI: https://doi.org/10.1037//0096-1523.7.5.1074

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. DOI: https://doi.org/10.1126/science.7777863

Thomas, D. C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., & Armstrong, B. G. 1985. The problem of multiple inference in studies designed to generate hypotheses. *American Journal of Epidemiology*, *122*(6), 1080–1095. DOI: https://doi.org/10.1093/oxfordjournals.aje.a114189

Tukey, J. W. 1953. *The problem of multiple comparisons* [Mimeographed notes]. Princeton, NJ: Princeton University.

Tukey, J. W. 1980. We need both exploratory and confirmatory. *The American Statistician, 34*(1), 23–25. DOI: https://doi.org/10.2307/2682991

van Heuven, V. J., & van Zanten, E. 2005. Speech rate as a secondary prosodic characteristic of polarity questions in three languages. *Speech Communication*, 47(1–2), 87–99. DOI: https://doi.org/10.1016/j.specom.2005.05.010

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. DOI: https://doi.org/10.1016/j.jml.2018.07.004

Vasishth, S., & Nicenboim, B. 2016. Statistical methods for linguistic research: Foundational ideas–Part I. *Language and Linguistics Compass*, *10*(8), 349–369. DOI: https://doi.org/10.1111/lnc3.12201

Vatikiotis-Bateson, E., Barbosa, A. V., & Best, C. T. 2014. Articulatory coordination of two vocal tracts. *Journal of Phonetics*, *44*, 167–181. DOI: https://doi.org/10.1016/j.wocn.2013.12.001

Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S* (4th Edition). Springer, New York. DOI: https://doi.org/10.1007/978-0-387-21706-2

von der Malsburg, T., & Angele, B. 2017. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of memory and language*, *94*, 119–133. DOI: https://doi.org/10.1016/j.jml.2016.10.003

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., & van Snellenberg, J. X. 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, *45*, S210–S221. DOI: https://doi.org/10.1016/j.neuroimage.2008.10.061

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. DOI: https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package = tidyverse.

Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. DOI: https://doi.org/10.1016/j.wocn.2018.03.002

Winter, B. 2011. Pseudoreplication in phonetic research. In: *Proceedings of the International Congress of Phonetic Science*, 2137–2140. Hong Kong.

Winter, B. 2014. Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967. DOI: https://doi.org/10.1002/bies.201400028

Winter, B. 2015. The other N: The role of repetitions and items in the design of phonetic experiments. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: The University of Glasgow.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1–2), 23–43. DOI: https://doi.org/10.1016/s0167-6393(98)00048-x

How to cite this article: Roettger, T. B. 2019 Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 1, pp. 1–27. DOI: https://doi.org/10.5334/labphon.147

Submitted: 20 March 2018 **Accepted:** 15 November 2018 **Published:** 04 January 2019

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS